International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 10 (2018) pp. 7256-7260
© Research India Publications. http://www.ripublication.com

On Improving the k-means Algorithm to Classify Unclassified Patterns

Mohamed M. Rizk[1], **Safar Mohamed Safar Alghamd**i[2]

[1]*Mathematics & Statistics Department, Faculty of Science, Taif University, Taif, Saudi Arabia.*
*Permanent Address***:** *Mathematics Department, Faculty of Science, Menoufia University, Shebin El-Kom, Egypt.*

[2]*Mathematics & Statistics Department, Faculty of Science, Taif University, Taif, Saudi Arabia.*

Abstract:

In this paper, the principal component analysis and rough sets to feature reduction in pattern recognition are used, which we applied this methods to get reduction of features in order to accelerate the k-means clustering after the relevant features selection, then the classification become more faster to classify any unknown pattern.

Keywords: Clustering, principal component analysis, rough sets, k-means.

INTRODUCTION

The dimension reduction of dataset is a one of fundamental steps in classifier design; it is needed when the dataset has a large number of features. Features may be irrelevant or relevant, there are two different manners to reduce the features; feature selection which considering a subset of the original feature and feature extraction which transforming the original features to extract a smaller amount of new features [1], [5]. In this paper, we will use principal components analysis [4], and rough sets [13-15], in the context of feature extraction and feature selection respectively.

The principle components analysis provides feature extraction and reduction optimal from the point of view of minimizing the reconstruction error, which we suppose that a large variance corresponds to useful information and little variance corresponds to useless information when using principle components analysis. Therefore, if the last few features have little variance we can discard them without losing too much information.

The rough set theory proposed by Pawlak in 1982 is an important tool to deal with imprecise, incomplete and inconsistent information [6-10]. In rough sets theory, the feature (attribute) reduction is one of the most important research contents. Most of feature reduction algorithm base on discernibility matrix [11-13], both the rows and columns of the matrix correspond to the objects. An element of the matrix is the set of all features that distinguish the corresponding object pairs, namely, the set consists of all features on which the corresponding two objects have distinct values. A rough set is based on its lower and upper approximations of a set, the approximation space and models of sets.

We know that the ability of collecting and storing large amounts of data around any field is a one of the main characterizations of this age, consequently, the processes of discovering patterns from this data need mathematical methods for removing irrelevant dimensions without affecting the accuracy of results. Also, classifications of this data are a main step of analysis in decision making and pattern recognition.

Recall that, principle component analysis being an unsupervised numeric approach, while rough set is a supervised symbolic approach, and principle component analysis is primarily a tool for analysing continuous data, and rough set is primarily a tool for analysing nominal data, that is; principle component analysis and rough set are fundamentally different in the way they analyses and manipulate data, but both techniques can also be used for analysing coarsely discrete data. Many datasets do contain features with a small number of discrete values, and for preprocessing such datasets both principle component analysis and rough set are potential candidates.

Clustering is the partition of a set into subsets so that the elements in each subset share some common treat, that is, it is a process of forming groups of objects, or clusters, in such a way that objects in one cluster are very similar and objects in different clusters are dissimilar. The k-means algorithm is the most widely used method for discovering clusters in data.

In this paper, we show how accelerating the k-means algorithm to classify unclassified patterns, which for large datasets the algorithm of k-means is slow in practice. The k-means algorithm based on n the number of objects (data points), k the number of clusters to be found, and i the number of iterations required, consequently the number of distance computations when we use the k-means algorithm is nki, Empirically, the number of iterations i grows with n, k and d the features of the objects (the dimensionality of the data), so we present a new algorithm based on the results of principal component analysis and rough sets theory to reduce the features d and to start the initial centers such that we partition the set of objects to sets and compute the centers for them to help us to suggest the number of clusters k, and the inequality Elkan [2] under some condition to reduce the number of iterations i is used, this in the case of unsupervised recognition.

ALGORITHMS

In this section, we suppose that the knowledge about a domain of recognition is represented by a limited size sample of n random d-dimensional patterns $\boldsymbol{x} \in R^d$ representing extracted object's features, and assume that an unlabeled training data set $D = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \cdots, \boldsymbol{x_n}\}$ can be represented as an $n \times d$ data pattern matrix $X = [\boldsymbol{x_1}, \boldsymbol{x_2}, \cdots, \boldsymbol{x_n}]^T$.

The training data set can characterized statistically by $d \times d$ dimensional covariance matrix Σ, then the eigenvalues of the covariance matrix Σ are computed, and they arranged in the decreasing order as following $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$ with the corresponding orthonormal eigenvectors $\boldsymbol{e_1}, \boldsymbol{e_2} \cdots, \boldsymbol{e_d}$, this gives us the components in order of significance, by using the definition selection method, we select the reduced dimension $p \leq d$ of a feature vector in the principal components space, let the set of reduce features denoted by R, i.e., R containing only the most dominant features.

Since the rough sets methods require that a processed data set contain discrete features, we discretize the patterns in R with respect to the condition features, apply rough sets in order to help us to suggest the number of clusters k by determinate the initial centers, consequently, we can partition the set of objects into sets (clusters), compute the centers c_i for $i = 1,2,\cdots,k$ and the distance $d(c_i, c_j)$ for all centers $c_i \neq c_j$, and find $\frac{1}{2} d(c_i, c_j)$.

Apply the inequality of Elkan [2] for a point x and two centers c_i, c_j, such that $c_i \neq c_j$ under the condition $d(c_i, c_j) \geq 2d(x, c_i)$, we obtain $d(x, c_j) \geq d(x, c_i)$, so it is not necessary to calculate $d(x, c_j)$, also; If any point is far away from a center, it is not necessary to calculate the exact distance between this point and the center in order to know that the point should not be assigned to this center. Conversely, if any point is much closer to one center than to any other, calculating exact distances is not necessary to know that the point should be assigned to the first center.

The inequality of Elkan [2] applies for every center different from c_i, so for most some distance calculations are avoided, even if others must be performed, consequently; it helps us to avoid the redundant distance calculations and accelerated the k-means algorithm. Finally, we can classify any unclassified pattern to the nearest cluster (nearest center) for it.

Algorithm 1:

If we have an unlabeled training data set $D = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \cdots, \boldsymbol{x_n}\}$ containing d-dimensional patterns, with real-valued features can be represented as an $n \times d$ data pattern matrix $X = [\boldsymbol{x_1}, \boldsymbol{x_2}, \cdots, \boldsymbol{x_n}]^T$.

1- For the matrix X subtract the mean from each of the data dimensions, and compute the covariance matrix Σ.

2- Compute the eigenvalues and corresponding eigenvectors for the covariance matrix Σ, and arrange their in a descending order.

3- Select the reduced dimension $p \leq d$ of a feature vector in the principal components space using the defined selection method, which may base on a judgement of the ordered values of computed eigenvalues, and denote the set of reduce features by R.

4- Discretize the patterns in R and use the rough sets to find the initial centers, which it is suggested the number of clusters k by partition the set of objects into sets (clusters) and compute the center c_i for $i = 1,2,\cdots,k$.

5- Compute the distance $d(c_i, c_j)$ for all centers $c_i \neq c_j$ and find $\frac{1}{2} d(c_i, c_j)$.

6- Apply the inequality of Elkan [2] to avoid the redundant distance calculations, we obtain $d(x, c_j) \geq d(x, c_i)$ for a point x and two centers c_i, c_j, such that $c_i \neq c_j$ under the condition $d(c_i, c_j) \geq 2d(x, c_i)$.

7- Classify any unclassified pattern to the nearest cluster for it.

EXPERIMENTAL RESULTS

Now we discuss the Iris data as an application example of our algorithm to explain how to apply the steps of algorithm, this data have been used as a benchmark for the linear discriminant analysis in Fisher (1936) [3], and it has been considered many times in the literature of clustering research.

Recall that, principle component analysis is an unsupervised method, and the data in this problem is taken labeled in order to show the accuracy of the present algorithm, and how it classifies the patterns before and after reducing the features. So the principle component analysis procedure is applied for an unlabeled training data set $D = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \cdots, \boldsymbol{x_{150}}\}$, such that we can be represented it as a (150×4) data pattern matrix $X = [\boldsymbol{x_1}, \boldsymbol{x_2}, \cdots, \boldsymbol{x_{150}}]^T$, where 4 is the length of the original pattern X (the condition features in rough set definition).

We obtain:

$Covariance\ matrix\ \Sigma =$

1.00671141	− 0.11010327	0.87760486	0.82344326
−0.11010327	1.00671141	− 0.42333835	− 0.358937
0.87760486	− 0.42333835	1.00671141	0.96921855
0.82344326	− 0.358937	0.96921855	1.00671141

$Eigenvectors\ (\boldsymbol{e_1}, \boldsymbol{e_2}, \boldsymbol{e_3}, \boldsymbol{e_4}) =$

0.52237162	− 0.37231836	− 0.72101681	0.26199559]
−0.26335492	− 0.92555649	0.24203288	− 0.12413481
0.58125401	− 0.02109478	0.14089226	− 0.80115427
0.56561105	− 0.06541577	0.6338014	0.52354627

$Eigenvalues\ in\ descending\ order\ (\lambda_1, \lambda_2, \lambda_3, \lambda_4) =$

2.93035378	0.92740362	0.14834223	0.02074601

From the values of eigenvalues we can reduce the $4-$dimensional feature space to a $2-$dimensional feature subspace (such that the two highest eigenvalues explain 95.8% of the total eigenvalues), by choosing the first two eigenvectors that corresponding the two highest eigenvalues to construct the following matrix:

0.52237162	-0.37231836
-0.26335492	-0.92555649
0.58125401	-0.02109478
0.56561105	-0.06541577

Then, the principal components as linear combinations of the original variables are obtained, and they are denoted by R. Discretize the patterns in R into four qualitative terms and use the rough sets, we find that the most of objects divided into three sets (clusters), so that we suggested the number of clusters $k = 3$, that is, we have three initial centers. Compute the center c_i for $i = 1,2,3$, the distance $d(c_i, c_j)$ for all centers $c_i \neq c_j$, and find $\frac{1}{2} d(c_i, c_j)$.

Apply the inequality of Elkan [2] for all objects in R (the objects either in the center c_i for $i = 1,2,3$ or not) and two centers c_i, c_j for $c_i \neq c_j$, which under the condition $d(c_i, c_j) \geq 2d(x, c_i)$ we obtain $d(x, c_j) \geq d(x, c_i)$, repeated steps 5 and 6 we obtain $k = 3$, therefore we avoid the redundant distance calculations and accelerate the k-means algorithm. The accuracy of approximation is 85.675%.

Now we discuss the following question in [14,15], what happens when the important information is hidden in features of small variation (this truth can be happen), that is, principle components analysis does not guarantee that selected first principal components, as a feature vector, will be adequate for classification. Nevertheless, the projection of high-dimensional patterns into lower dimensional orthogonal principal component feature vectors might help to provide better classification for some data types, one of the possibilities to improve that to apply rough set theory. Note that, on the other hand, many researchers' shows that real datasets often have the characteristics that make principle components analysis suitable for data reduction, i.e., the important information in a dataset to be concentrated in the first few principal components.

The following algorithm can be proposed for the principle components analysis and rough sets.

Assume that we have as above an unlabeled training data set $D = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \cdots, \boldsymbol{x_n}\}$, and it is represented as a $n \times d$ data pattern matrix $X = [\boldsymbol{x_1}, \boldsymbol{x_2}, \cdots, \boldsymbol{x_n}]^T$. The covariance matrix Σ $d \times d$ dimensional is founded, then the eigenvalues of Σ are computed, and they arranged in the decreasing order as following $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$ with the corresponding orthonormal eigenvectors $\boldsymbol{e_1}, \boldsymbol{e_2} \cdots, \boldsymbol{e_d}$, this gives us the components in order of significance, by using the definition selection method. The reduced dimension $p \leq d$ of a feature vector in the principal components space is selected, and let the set of reduce features denoted by R_1, i.e., R_1 containing only the most dominant features. The patterns in D with respect to the condition features are discretized, then rough sets to reduce the features are applied, the reduct is computed, and denoted by R_2. Put $R_A = R_1 \cup R_2$, i.e., R_A is the union of features in R_1 and R_2. Apply rough set again in order to help us to suggest the number of clusters k by determinate the initial centers, consequently, we can partition the set of objects into sets (clusters) and compute the center c_i for $i = 1,2, \cdots, k$.

Compute the distance $d(c_i, c_j)$ for all centers $c_i \neq c_j$ and find $\frac{1}{2} d(c_i, c_j)$. Apply the inequality of Elkan [2] for all objects in R_A (the objects either in the center c_i for $i = 1,2, \cdots, k$ or not) and two centers c_i, c_j for $c_i \neq c_j$, which under the condition $d(c_i, c_j) \geq 2d(x, c_i)$ we obtain $d(x, c_j) \geq d(x, c_i)$, it helps us to avoid the redundant distance calculations and accelerated the k-means algorithm. Finally, we can classify any unclassified pattern to the nearest cluster (nearest center) for it.

Algorithm 2:

Suppose as in algorithm 1 that we have an unlabeled training data set $D = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \cdots, \boldsymbol{x_n}\}$ containing d-dimensional patterns with real-valued features can be represented as an $n \times d$ data pattern matrix $X = [\boldsymbol{x_1}, \boldsymbol{x_2}, \cdots, \boldsymbol{x_n}]^T$.

1- For the matrix X subtract the mean from each of the data dimensions, and compute the covariance matrix Σ.

2- Compute the eigenvalues and corresponding eigenvectors for the covariance matrix Σ, and arrange their in a descending order.

3- Select the reduced dimension $p \leq d$ of a feature vector in the principal components space using the defined selection method, which may base on a judgement of the ordered values of computed eigenvalues denoted by R_1.

4- Discretize the patterns in X with the condition features, and compute the reduct set and denoted by R_2.

5- Take the union of features in steps 4 and 5, that is, $R_A = R_1 \cup R_2$.

6- Rough sets used again to find the initial centers, which it is suggested the number of clusters k by partition the set of objects into sets (clusters) and compute the center c_i for $i = 1,2, \cdots, k$.

7- Compute the distance $d(c_i, c_j)$ for all centers $c_i \neq c_j$ and find $\frac{1}{2} d(c_i, c_j)$.

8- Apply the inequality of Elkan [2] to avoid the redundant distance calculations, we obtain $d(x, c_j) \geq d(x, c_i)$ for a point x and two centers c_i, c_j, such that $c_i \neq c_j$ under the condition $d(c_i, c_j) \geq 2d(x, c_i)$.

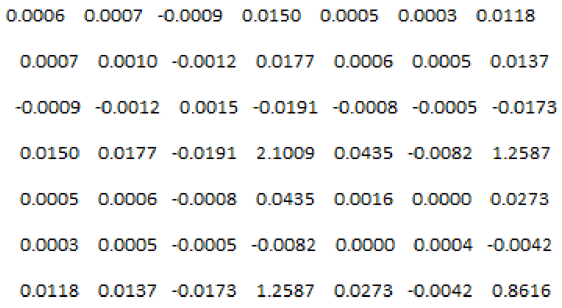
9- Classify any unclassified pattern to the nearest cluster for it.

EXPERIMENTAL RESULTS

Suppose that we have a practical problem from experimental results in biochemistry as typical data, a small data set taken from the literature is used (see table (20) [17]), this data concern modeling of the energy for unfolding of a protein (tryptophan synthase alpha unit of the bacteriophage $T4$ lysozome), where 19 coded amino acids (AAs) were each introduced. The AAs are described in terms of seven features.

Note that, we take the data in this problem labeled also as in algorithm 1 in order to show the accuracy of the above algorithm, and how it classifies the patterns before and after reducing the features. So the principle component analysis procedure is applied for an unlabeled training data set $D = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \cdots, \boldsymbol{x_{19}}\}$, and D is represented as an (19×7) data pattern matrix $X = [\boldsymbol{x_1}, \boldsymbol{x_2}, \cdots, \boldsymbol{x_{19}}]^T$, where 7 is the length of the original pattern X, i.e., the only condition features. We obtain:

$Covariance\ matrix\ \Sigma =$



0.0006	0.0007	-0.0009	0.0150	0.0005	0.0003	0.0118
0.0007	0.0010	-0.0012	0.0177	0.0006	0.0005	0.0137
-0.0009	-0.0012	0.0015	-0.0191	-0.0008	-0.0005	-0.0173
0.0150	0.0177	-0.0191	2.1009	0.0435	-0.0082	1.2587
0.0005	0.0006	-0.0008	0.0435	0.0016	0.0000	0.0273
0.0003	0.0005	-0.0005	-0.0082	0.0000	0.0004	-0.0042
0.0118	0.0137	-0.0173	1.2587	0.0273	-0.0042	0.8616

$Eigenvalues\ in\ descending\ order\ (\lambda_1, \lambda_2, \cdots, \lambda_7) = 1.0\text{e} + 03$

$\lambda_1 = 2.885656290480553$ $\lambda_2 = 0.07871244417166$

$\lambda_3 = 0.002594509337186$ $\lambda_4 = 0.000623368281622$

$\lambda_5 = 0.000030665023762$ $\lambda_6 = 0.000021363547674$

$\lambda_7 = 0.000001551233566$

So, we have 7 eigenvalues and corresponding 7 eigenvectors, from the values of eigenvalues we can reduce the $7-$dimensional feature space to a $2-$dimensional feature subspace (such that the two highest eigenvalues explain 99.89% of the total eigenvalues), by choosing the first two eigenvectors that corresponding the two highest eigenvalues. Then, the principal components as linear combinations of the original variables are obtained, and they are denoted by R_1 as a matrix (7×2).

Discretize the patterns for all eigenvectors into four qualitative terms (the condition features are coded into four qualitative terms). Use the rough set, and compute the reduct set R_2 we find that $R_2 = \{\boldsymbol{e_2}, \boldsymbol{e_5}\}$. Then $R_A = R_1 \cup R_2 = \{\boldsymbol{e_1}, \boldsymbol{e_2}, \boldsymbol{e_5}\}$, now used rough sets again to find the initial centers we find that the most of objects divided into two sets (clusters), so that we suggested the number of clusters $k = 2$, that is, we have two initial centers. Compute the center c_i for $i = 1,2$, and find $\frac{1}{2} d(c_1, c_2)$. Apply the inequality of Elkan [2] for all objects in R_A and two centers c_1, c_2 for $c_i \neq c_j$, which under the condition $d(c_1, c_2) \geq 2d(x, c_1)$ we obtain $d(x, c_2) \geq d(x, c_1)$, repeated steps 7 and 8 we obtain $k = 3$, therefore we avoid the redundant distance calculations and accelerate the k-means algorithm. The accuracy of this approximation is 86.875%.

CONCLUSION

In this paper, new algorithms based on principal component analysis and rough sets methods to reduce the features in the case of unsupervised recognition are found; consequently we accelerated the k-means algorithm in this case, and explained the new algorithms by some examples.

REFERENCE

[1] Cios K., Pedrycz W., and .Winiarski R.W. (1998) Data Mining Methods in Knowledge Discovery. Boston, Dordrecht, London: Kluwer Academic Publishers

[2] Elkan, Charles. Using the triangle inequality to accelerate k-means. In Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), August 21-24, 2003, Washington, DC, USA, pp. 147–153

[3] Fisher, R. A. (1936) The use of multiple measurements in taxonomic problems. *Ann. Eugen.*, **7**, 179–188.

[4] Jolliffe, I. T. (2002) Principal component analysis, New York, Springer-Verlag.

[5] Kittler, J., Feature selection and extraction. In Young and Fu (Eds.), (1986) Handbook of pattern recognition and image processing, pp203-217, New York: Academic Press.

[6] Pawlak Z, Rough sets [J]. (1982) International Journal of Computer and Information Sciences, 11 (5): 341-356.

[7] Pawlak, Z., (1991) Rough sets, theoretical aspects of reasoning about data, Kluwer Academic Publisher.

[8] Pawlak, Z., A. Skowron, A., (2007) Rough sets: some extensions, Inform. Sci., 177 (1) 28–40.

[9] Polkowski, L., Skowron, A. (Eds.), (1998) Rough Sets and Current Trends in Computing, vol. 1424, Springer, Berlin.

[10] Polkowski, L., Skowron, A. (Eds.), (1998) Rough Sets in Knowledge Discovery, vol. 1, Physica-Verlag, Berlin.

[11] Skowron, A., Rauszer, C., (1992) The discernibility matrices and functions in information systems, in: R. Slowiński (Ed.), Intelligent Decision Support, Handbook of Applications and Advances of the Rough Sets Theory, Kluwer. Dordrecht.

[12] Skowron, A., Stepaniuk, J., (1996) Tolerance approximation spaces, Fundamental Informatcae. 27: 245-253.

[13] Slowinski, R., Vanderpooten, D., (2000) A generalized definition of rough approximations based on similarity,

IEEE Trans. Knowledge Data Eng., 12 (2) 331–336.

[14] Swiniarski, R, (2001) Rough Sets Methods in Feature Reduction and classification. Int. J. Appl. Math. Comput. Sci., Vol.11, No.3, 565-582.

[15] Swiniarski, R. Skowron, A., (2003) Rough Sets Methods in Feature Selection and Recognition. Pattern Recognition Letters. 24(6). 833-849.

[16] Theodoridis S., Koutroumbas K. (2009) Pattern recognition, 4th ed., Academic Press.

[17] Walczka B. and Massart D.L., Tutorial rough sets theory, chemometrics and intelligent laboratory systems, 1999, Vol. 47, 1-16.

[18] Zdzislaw Pawlak, (2002) Rough set theory and its applications, Journal of Telecommunications and Information Technology 3, 7-10.